SYSTEM, METHOD, AND SERVICE FOR COLLABORATIVE FOCUSED CRAWLING OF DOCUMENTS ON A NETWORK

ABSTRACT OF THE INVENTION

A collaborative focused crawler crawls documents on a network locating documents that match multiple focus topics. The collaborative crawler comprises a fetcher and a focus engine. The fetcher prioritizes which documents to crawl based on a set of rules, obtains documents from the network, and outputs crawled documents to the focus engine. The focus engine determines whether a fetched document is relevant to any of the multiple focus topics. The focus engine determines whether fetched documents are disallowed. If a fetched document is disallowed, the present system may place the URL for that web document in a blacklist, a list of URLs that may not be crawled. URLs may be disallowed if they match a disallowed topic or if they fail a set of rules designed for a web space focus, for example, domain rules, IP address rules, and prefix rules.